Revolutions Under Quarantine: Algorithmic Containment, Micro-Elites, and the Two-Front Struggle

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

September 16, 2025

This paper examines "constant revolution" when it is driven not by campuses or newsrooms but by tiny, siloed micro-elites on social platforms—people whom algorithms do **not** sponsor into the public square but **sort and suppress** into quarantined niches. In this regime, platforms act as risk managers: they dampen cross-silo reach (downranking, de-recommendation, demonetization), allow local visibility to persist, and thereby keep high-energy discourse **contained**. The result is a two-front struggle. Micro-elites fight upstream against the platform's containment architecture while also contesting legacy elites and government power. We model an adversarial **I–G_c–E–P** system: micro-elites (I), a containment-oriented algorithmic gatekeeper (G_c), legacy elites and state actors (E), and the broader public (P). We predict symbolic turbulence within silos, short attention half-lives outside them, and rare but sharp spills when containment misfires or elites split. We propose measurable signatures—reach suppression ratios, cross-silo path entropy, pickup-to-policy translation rates—and outline falsifiers that would disprove the thesis. Finally, we sketch both strategy (poly-platform routing, "proof-of-outcome" activism) and design options (auditable ranking, pluralized cues) aimed at reducing brittleness while reconnecting attention with real-world results.

Keywords: algorithmic containment, silo quarantine, downranking, de-recommendation, demonetization, shadow suppression, micro-elites, audience capture, visibility tax, cross-silo spillover, platform hopping, dogwhistle adaptation, legacy elites, legitimacy, attention half-life, signal brokerage, proof-of-outcome activism, governance, falsifiability.

A collaboration with GPT-5. 49 pages. CC 4.0.

Outline

1. Problem and Reframe
   1.1 From institution-led to platform-quarantined “revolutions”
   1.2 Actors and notation: I–G_c–E–P
   1.3 Scope conditions: high online energy, low offline conversion
2. The Containment Architecture
   2.1 Mechanisms: downranking, de-recommendation, demonetization, policy strikes
   2.2 Silo formation and maintenance: similarity metrics, risk flags, feedback loops
   2.3 The platform’s objective: minimize systemic risk, preserve engagement locally
3. Adversarial I–G_c–E–P Model
   3.1 State variables: local reach, cross-silo path entropy, pickup rate, policy translation
   3.2 Dynamics: containment equilibrium, leakage, and cascade conditions
   3.3 Comparative statics: moderation shifts, regulatory shocks, elite splits
4. Outcomes and Phase Paths
   4.1 Symbolic turbulence with low material change (visibility debt)
   4.2 Platform Thermidor: tightening rules → durable quarantine
   4.3 Misfire moments: leak-through and rapid realignment
   4.4 Burnout and churn within micro-elites
5. Measurement and Indicators
   5.1 Reach-suppression ratio (impressions per follower vs baseline)
   5.2 Attention half-life outside the silo
   5.3 Cross-venue pickup and policy-translation ratios
   5.4 Search/sentiment deltas among non-followers
6. Strategies for Siloed Actors
   6.1 Poly-platform routing and redundancy
   6.2 Link laundering vs credibility loss: a tradeoff
   6.3 “Proof-of-outcome” activism: FOIA wins, service delivery, verifiable impacts
   6.4 Bridge nodes: cultivating cross-silo intermediaries
7. Governance and Design Options
   7.1 Auditable ranking signals and contestable downranking
   7.2 Frictions that reward verification over virality
   7.3 Pluralizing cues: diversified recommendation neighborhoods
   7.4 Safe harbors for cross-silo deliberation

8. Tests, Competing Explanations, and Falsifiers
   8.1 What empirical patterns would disconfirm “quarantine”
   8.2 Distinguishing resource poverty from algorithmic suppression

9. Limitations and Ethical Considerations
   9.1 Privacy, measurement bias, and adversarial inference
   9.2 Risks of chilling legitimate dissent

10. Appendices (Operationalization)
   A. Data schema and metrics cookbook
   B. Simulation sketch for I–G_c–E–P dynamics
   C. Survey modules for private/public belief gaps

1. Problem and Reframe

1.1 From institution-led to platform-quarantined “revolutions”

Classical accounts assume elite institutions (universities, parties, newsrooms, unions) birth and broadcast revolutionary frames, then mobilize the public. Our setting is inverted. A small fraction of high-output accounts—algorithmically clustered and partially suppressed—generate near-continuous rhetorical heat **inside** platform silos. Platforms act less like megaphones and more like **containment layers**: they allow local visibility but reduce cross-silo spread (downranking, de-recommendation, demonetization). The result is a **quarantined revolution**: high symbolic velocity, low social penetration.

Key contrasts:

- Gatekeeping: editorial committees → ranking functions with risk flags.
- Scale: broadcast reach → silo reach capped by suppression and similarity scores.
- Feedback: institutional outcomes (laws, budgets) → engagement loops (likes, replies) with weak links to policy.
- Time profile: long campaigns → flash-to-ash cycles with short attention half-lives outside the silo.

Net effect: participants experience perpetual upheaval; the median citizen sees little change unless containment misfires or legacy elites pick up and launder the signal.

1.2 Actors and notation: I–G_c–E–P

We model four interacting nodes and a few measurable links:

- **I (micro-elites):** small, high-activity creators/organizers operating inside silos.
    - State: content cadence c_t, network density d_t, bridge nodes b_t.
    - Observable: local impressions R_local, follower growth g_t, external pickup kappa_t.
- **G_c (containment gatekeeper):** platform algorithms and policies oriented to risk minimization over open diffusion.
    - Levers: suppression s_t (0..1), recommendation exposure q_t, monetization m_t, strike rate sigma_t.
    - Goal (stylized): minimize systemic-risk proxy while preserving aggregate engagement.
- **E (legacy elites):** government, major media, large NGOs/corporates.
    - Behavior: selective pickup L_t of silo narratives; conversion to agenda items A_t; policy throughput tau_t.
- **P (public):** median citizen and broad audiences outside the silo.
    - State: awareness W_t, salience S_t, behavioral uptake U_t (donations, attendance, votes, purchases).

Minimal link variables (ASCII-safe):

- Cross-silo reach: R_cross_t = R_0 * (1 - s_t) * x_t, where x_t captures bridge quality and topical resonance.
- Legacy pickup: L_t = f(R_cross_t, elite incentives, news cycle constraints).
- Policy translation: tau_t = g(L_t, institutional capacity, veto points).
- Attention half-life outside the silo: hl_t = h(content form, controversy level, counter-messaging).

Interpretation: with s_t high, R_cross_t and downstream tau_t stay low even if I generates large R_local. “Revolution” remains symbolic unless (i) s_t drops, (ii) x_t rises (better bridges), or (iii) E splits and amplifies L_t.

1.3 Scope conditions: high online energy, low offline conversion

Our framework applies under the following conditions:

- **Platform dominance of discovery:** a large share of agenda discovery for P flows through ranked/recommended feeds; search and direct navigation play secondary roles.
- **Active containment:** sustained downranking/de-recommendation (s_t high) for the focal topics or accounts; monetization limits that raise a "visibility tax."
- **Sparse offline infrastructure:** limited unions, churches, clubs, or field organizations able to convert attention into durable coalitions; event logistics and service provision are thin.
- **High rhetorical velocity, low material throughput:** abundant memes, petitions, boycotts; few budget lines changed, laws passed, or services delivered.
- **Cue dependence:** P updates beliefs/actions primarily via elite cues (E) rather than first-hand experience or peer networks.

Boundary cases (where the model may not fit or predicts transitions):

- **Open amplification windows:** sudden policy shifts, regulatory pressure, or product changes reduce s_t → temporary leak-through.
- **Bridge proliferation:** growth in b_t (journalists, cross-community creators, NGOs) increases x_t → higher L_t without s_t changing.
- **Crisis shocks:** exogenous X_t (scandal, economic shock, disaster) forces E to pick up silo narratives → tau_t rises quickly.
- **Strong offline capacity:** when local organizations exist to capture attention and deliver outcomes, U_t can grow despite high s_t.

Empirical fingerprint of the scope: R_local ≫ R_cross; hl_t outside the silo is short; L_t and tau_t are intermittent and clustered around misfires or elite splits.

2. The Containment Architecture

2.1 Mechanisms: downranking, de-recommendation, demonetization, policy strikes

We model platform moderation and ranking as a multi-stage pipeline, where containment is achieved mostly by *distribution throttles* rather than removals.

Pipeline sketch (conceptual):
Candidates → Safety filters → Ranking → Integrity checks → Monetization gate → Delivery

- **Downranking (distribution dampers).**
  Soft suppression applied inside ranking rather than outright removal. If a baseline exposure score is R_0, the delivered exposure is

- R = R_0 * (1 - s_t) * q_t

where s_t in [0,1] is a suppression factor (policy/quality), and q_t in (0,1+] captures quality boosts/deboosts (e.g., clickbait or trusted-source signals). Containment operates by raising s_t and lowering q_t for targeted content, shrinking cross-silo reach without notifying creators.

- **De-recommendation (candidate pruning).**
  Content is eligible for feeds of *followers* but pruned from *non-followers* via candidate generators. This removes the primary growth channel. Operationally: set recommendability flag rec=0 for certain topics/accounts; the content remains "public," but discovery decays.

- **Demonetization (visibility tax).**
  Ads/revenue sharing are withheld (m_t -> 0). Since creators optimize for sustainable production, expected output c_{t+1} falls with E[rev_t]. Demonitized content often correlates with lower recommendation exposure as an integrity heuristic.

- **Policy strikes and enforcement ladders.**
  Violations increment a strike counter sigma_t. Thresholds trigger temporary limits (reduced upload frequency, comment freezes), persistent downranking, or suspensions. Even after the strike window expires, trust scores may recover slowly (hysteresis), maintaining containment.

How this feels on the ground: impressions flatten relative to follower count; virality becomes rare; growth stalls; no explicit "ban" exists to contest. The revolution is *legible* to participants but *invisible* to most outsiders.

2.2 Silo formation and maintenance: similarity metrics, risk flags, feedback loops

Containment does not merely shrink reach; it *shapes* network topology.

- **Similarity metrics and neighborhood curation.**
  Candidate generation relies on embeddings and co-engagement graphs. Let sim(u,v) be cosine similarity in content/interaction space. Silo cores emerge when:

- P(exposure u→v) ∝ sim(u,v) * (1 - s_t(u,v))

High similarity plus targeted suppression of cross-cutting edges reduces *path entropy* between communities, keeping discourse local.

- **Risk flags and topic buckets.**
  Content receives latent risk labels (e.g., "heated politics," "health claims," "coordinated harassment"). These flags add dampers at multiple stages:
    - rec=0 outside followers,
    - stricter spam/duplication thresholds,
    - higher friction for link-outs or off-platform calls to action.
- **Feedback loops (self-reinforcing silos).**

1. **Engagement loop:** local engagement is rewarded, encouraging insiders to optimize for in-silo norms (jargon, priors), which lowers cross-silo similarity further.

2. **Suppression loop:** flagged speech adapts via euphemisms/dogwhistles; detection models update; containment persists.

3. **Bridge attrition:** cross-silo "bridge nodes" see lower ROI (fewer impressions, higher moderation risk), reducing the very edges that could ventilate the silo.

- **Adversarial adaptation.**
  Creators respond with lexical drift, format changes (images/video/subtitles), or platform hopping. Platforms counter with updated detectors and quality heuristics. The equilibrium is *containment with leakage*: enough local visibility to retain users, insufficient cross-silo diffusion to reshape the public agenda.

Measurable fingerprint (outside audits):
R_local >> R_cross, declining ratio of non-follower impressions per post, rising average homophily in commenter networks, shortening attention half-life among non-followers, and a falling share of external search traffic.

2.3 The platform's objective: minimize systemic risk, preserve engagement locally

Platforms face a multi-objective problem:

- **Primary constraint:** reduce tail-risk events (societal harm, regulatory crises, advertiser boycotts).
- **Commercial imperative:** maintain session length, creator retention, and ad safety.

- **Operational feasibility:** prefer scalable, automated controls over case-by-case editorial adjudication.

We capture this with a stylized objective:

Maximize $E[Engagement_t] - \lambda * Risk_t$

Subject to $Moderation_cost_t \leq M_t$

where $Risk_t$ aggregates regulatory, reputational, and real-world spillover risks; λ is a time-varying penalty weight (rises after scandals/elections).

Why containment (vs removal) is attractive:

- **Risk minimization:** lowering R_{cross} shrinks the chance of broad cascades without the PR cost of bans.
- **Local engagement preservation:** in-silo activity keeps users busy, limiting churn and preserving ad inventory.
- **Scalability:** downranking and de-recommendation are model-driven and adjustable by policy toggles, requiring fewer human reviews.

Implications for our model (I–G_c–E–P):

- With λ high, the platform raises s_t on flagged clusters, depressing R_{cross} even for accurate or constructive content that shares features with risky discourse.
- Legacy elites (E) encounter fewer organic signals from silos; pickup L_t becomes rarer and more correlated with exogenous shocks.
- Public awareness W_t depends more on E-mediated cues than on direct platform diffusion, keeping offline conversion low unless containment misfires or bridges proliferate.

Design tension:
Containment can reduce systemic risk yet also *distort public learning*, privileging safe consensus over empirically valid but contentious claims. Our later sections propose governance levers (auditable ranking signals, contestable downranking, outcome-linked credibility) to narrow this gap without reopening the floodgates to genuine harm.

3. Adversarial I–G_c–E–P Model

3.1 State variables: local reach, cross-silo path entropy, pickup rate, policy translation

We track four core observables (per time step t), plus a few control parameters.

Core state

- R_local_t : impressions within the silo (followers + near-neighbors).
- R_cross_t : impressions among non-followers outside the silo.
- H_path_t : cross-silo path entropy (diversity of short paths from the silo to other communities in the interaction/reco graph).
- L_t : legacy pickup rate (probability or share of cycles in which E "launders" the signal into mainstream coverage).
- tau_t : policy translation rate (fraction of picked-up items converting to agenda entries, budget lines, or rules).

Controls (stylized, measurable)

- s_t in [0,1] : suppression factor (downranking, de-recommendation intensity).
- x_t >= 0 : bridge quality (density and credibility of cross-silo bridge nodes).
- q_t : content quality/trust features (source diversity, factuality).
- lambda_t >= 0 : platform risk-aversion weight (higher after scandals/regulatory heat).
- kappa_t : elite "news hunger"/incentive to pick up outsider narratives.
- v_t in [0,1] : veto friction in institutions (committee gates, legal constraints).
- X_t : exogenous shock indicator (scandals, crises).

Measurement notes

- H_path_t can be estimated via random-walk path entropy over the cross-silo neighbor distribution; normalize by log K (K reachable communities) to get h_t in [0,1].
- R_cross_t is proxied by non-follower impressions or "first-seen" exposures among out-groups.

Link functions (minimal, ASCII-safe)

(1) R_cross_t = R_potential_t * (1 - s_t) * x_t * f(h_t), with f'(h) > 0, 0 ≤ h_t ≤ 1

(2) L_t = sigmoid(a0 + a1*R_cross_t + a2*kappa_t + a3*X_t - a4*lambda_t)

(3) tau_t = L_t * (1 - v_t) * mu

Here mu in (0,1] is a conversion efficiency; sigmoid(z) = 1/(1+e^-z).

3.2 Dynamics: containment equilibrium, leakage, and cascade conditions

State updates (one simple specification)

(4) h_{t+1} = h_t + eta1*bridge_invest_t - eta2*silo_reinforce_t - eta3*s_t

(5) s_{t+1} = s_t + gamma1*Risk_t - gamma2*Audit_t - gamma3*Compete_t

(6) R_potential_{t+1} = R_potential_t * (1 + rho_local*Reward_t - delta*Burnout_t)

- bridge_invest_t includes creator outreach, cross-community collaborations, journalist ties.
- silo_reinforce_t is homophilic optimization (in-silo jargon, outgroup dunking).
- Risk_t aggregates flagged-topic volume and prior spillovers; Audit_t covers transparency requirements, appeals; Compete_t captures platform-competition pressure to relax suppression.
- Reward_t reflects local engagement payouts; Burnout_t captures demonetization, strikes, and opportunity costs.

Containment equilibrium (CE)

A CE obtains when R_cross_t stays below a spill threshold Theta and tau_t ~ 0 despite high R_local_t.

Conditions (informal):

- High, persistent s_t (distribution throttles) and low h_t (thin bridges).
- lambda_t elevated (risk-averse stance), keeping L_t low even for valid claims.
- v_t high (strong veto points), so even picked-up narratives rarely translate.

Empirical fingerprint:

- R_local_t >> R_cross_t, short attention half-life outside the silo, sporadic and non-cumulative mainstream coverage.

Leakage regime (LR)

Leakage occurs when any of the following raise R_cross_t or L_t above CE levels:

- **Suppression relief:** s_t drops (policy change, enforcement fatigue, successful appeals).
- **Bridge expansion:** h_t rises via more/stronger cross-silo intermediaries (x_t up).
- **Shock-driven receptivity:** X_t spikes, temporarily lifting kappa_t (news hunger) or lowering editorial thresholds in E.

Cascade condition (CC)

Define a coarse "reproduction number" for issue spread across E and P:

(7) R_eff = beta * R_cross_t * (kappa_t) * (1 - v_t) * g_feedback

- beta maps exposure to belief/agenda adoption outside the silo.
- g_feedback ≥ 1 captures feedback loops (media–policy–search–social).

A **cascade** is likely when R_eff > 1 for enough consecutive steps that E's endorsement seeds additional E/P endorsements (branching-like growth). When R_eff ≤ 1, surges decay (flash-to-ash).

Overshoot/backlash

If leakage coincides with high lambda_t (advertiser/regulator sensitivity), platforms can snap back (raise s_t), producing a visible spike then long trough ("platform Thermidor").

3.3 Comparative statics: moderation shifts, regulatory shocks, elite splits

We examine how exogenous parameter moves shift equilibria and observables.

A. Moderation shifts (platform policy toggles)

- ds > 0 (tightening): lowers R_cross_t via (1); over time, reduces h_t via (4) as bridges atrophy; increases "visibility debt" (online heat without offline payoff). Predict: lower non-follower impressions/post, higher homophily, longer CE spells.

- ds < 0 (loosening): raises R_cross_t; if R_eff crosses 1, expect short-run cascades; medium run, lambda_t may rise from perceived risk, partially undoing the loosening.

B. Regulatory shocks (penalties, transparency mandates, competition policy)

- **Penalty/PR shock:** increases lambda_t → L_t falls by (2); platforms raise s_t by (5). Predict: broader containment, slower cross-silo growth.
- **Transparency/right-to-appeal:** increases Audit_t → lowers s_t by (5); if bridge actors exist, h_t rises and leak-through becomes more frequent but still contingent on kappa_t.
- **Interoperability/competition:** raises Compete_t → lowers s_t; also raises effective x_t via poly-platform routing, making LR more common.

C. Elite splits (E loses unanimity)

- When E fragments (e.g., partisan media divergence, intra-institution dissent), the baseline threshold for L_t drops:
- L_t' = sigmoid(a0' + a1*R_cross_t + ...), with a0' > a0

Predict: for the same R_cross_t, more pickup events; if v_t also falls (aligned control of veto points), tau_t rises sharply. This can generate cascades without any change in s_t (containment intact, but E amplifies anyway).

D. Strengthening offline capacity (not a parameter shock, but a design lever)

- Investments that raise conversion efficiency mu or lower v_t increase tau_t directly (3), turning small L_t into visible outcomes. Predict: same online metrics, more offline wins (FOIA yields, local services, rule changes).

What would falsify the model

- Sustained periods where R_local_t ~ R_cross_t despite high s_t and low h_t.
- High tau_t without increases in L_t (policy changes happening absent elite pickup or strong offline capacity).
- No correlation between lambda_t shocks and subsequent s_t moves (5).

Analyst's checklist (operational)

- Track R_cross_t / R_local_t, normalized by topic and seasonality.
- Estimate h_t from cross-community exposure distributions; monitor trend.

- Log L_t via media coverage and agenda mentions; pair with tau_t (legislation, budgets, enforcement actions).
- Compute R_eff using calibrated beta, kappa_t, and v_t; watch crossings above 1 as early-warning for cascades.

4. Outcomes and Phase Paths

4.1 Symbolic turbulence with low material change (visibility debt)

Definition. *Visibility debt* is the gap between online salience and offline consequence: large R_local_t, noisy cycles inside the silo, yet persistently low L_t and tau_t. The movement "owes" reality outcomes it hasn't delivered.

Mechanism. With high s_t and low h_t, most energy recirculates locally. Engagement rewards in-silo fluency and ever-sharper rhetoric. Outside the silo, attention half-life is short; elites see little that is both reputationally safe and agenda-worthy.

Diagnostics.

- R_local_t >> R_cross_t; non-follower impressions/post fall over time.
- Coverage spikes without follow-on agenda mentions: L_t intermittent, tau_t ≈ 0.
- Search and donations from non-followers decay quickly after each "surge."

Predictions.

- Increasing frequency of "announcement" moments with few concrete wins.
- Frame drift toward symbolic victories (language, badges) over budget lines, laws, or services.

Tactics to reduce the debt.

- Reallocate effort to bridge building (x_t ↑, h_t ↑) and *proof-of-outcome* actions (FOIA wins, service delivery) that translate even with small L_t.

4.2 Platform Thermidor: tightening rules → durable quarantine

Definition. *Platform Thermidor* is the routinization of dissent into durable containment: rule updates and classifier retrains that harden s_t and depress h_t for long periods.

Triggers. Advertiser pressure, regulatory heat, scandal cycles (lambda_t ↑). Platforms respond with:

- broader risk buckets (topic-level dampers),
- lower recommendability (rec=0) for clusters,
- stricter monetization safety (effective m_t → 0).

Dynamics.

- **Hysteresis:** even after the headline crisis passes, trust and quality scores recover slowly; s_t remains high.
- **Topology effects:** bridge nodes suffer worst ROI and attrition; h_t trends down as cross-silo edges die.

Diagnostics.

- Structural drop in R_cross_t / R_local_t coincident with policy changes.
- Persistent demonetization flags; appeal win rates low.
- Rising homophily and decreasing path entropy in exposure graphs.

Predictions.

- Longer intervals between leak-through events; fewer cross-silo “first impressions.”
- Narrative radicalization inside the silo (to keep local engagement) paired with declining mainstream relevance.

Design levers (platform/governance).

- Contestable downranking with auditable reasons.
- Narrow, time-bounded risk buckets; automatic review/decay of s_t.
- Interop/competition pressures (Compete_t ↑) to check permanent quarantine.

4.3 Misfire moments: leak-through and rapid realignment

Definition. *Misfire moments* are short windows when containment fails (policy toggle, enforcement fatigue, algorithmic change, or exogenous shock X_t) and R_cross_t briefly jumps, raising L_t. If elite incentives (kappa_t) align and veto friction (v_t) is low, tau_t can spike.

Pathways.

1. **Algorithmic slip:** a recommender refresh lowers s_t unintentionally; bridge content rides generalized boosts.
2. **Shock-driven news hunger:** crises force E to relax thresholds; outsider material clears the bar.
3. **Elite splits:** factional cues in E make pickup reputationally safer.

Cascade test. If R_eff = beta * R_cross_t * kappa_t * (1 - v_t) * g_feedback > 1 for several steps, endorsement chains propagate beyond the silo.

Diagnostics (real time).

- Sudden rise in non-follower impressions and search queries among out-groups.
- Cross-venue pickup (local TV, mainstream print) within 24–72 hours.
- Policy hooks appear: hearings scheduled, RFPs drafted, budget riders floated.

Predictions.

- **Overshoot–snapback:** post-leak, lambda_t spikes → platform raises s_t harder (Thermidor snapback).
- **Path dependence:** if early outcomes materialize (tau_t > 0), later leak-through becomes easier (credibility cues).

Tactics for actors.

- Pre-stage “shovel-ready” artifacts (datasets, draft rules, coalitions) so that any leak converts to outcomes fast.
- Use bridge nodes with mixed reputations across communities to stabilize L_t before snapback.

4.4 Burnout and churn within micro-elites

Mechanism. High cadence under suppression (low m_t, frequent strikes sigma_t, flat R_cross_t) depletes creator slack. Audience capture pressures escalate rhetoric; enforcement risk rises; accounts cycle out.

Symptoms.

- Falling output c_t; longer gaps between posts; more platform hopping.
- Escalation to formatting "workarounds" (coded language, image text), then diminishing returns.
- Internal schisms over purity vs bridge-building; fragmentation of coalitions.

Costs to the system.

- Loss of bridge capacity as moderate, credibility-rich actors exit.
- Increased variance in content quality; higher false-positive risk for moderation.
- Deeper visibility debt as remaining actors optimize for in-silo validation only.

Metrics to watch.

- Creator survival curves (time-to-churn) before and after policy shifts.
- Share of content from "bridge-capable" accounts; ratio of collaborative to insular posts.
- Burnout proxies: demonetization duration, appeal latency, strike frequency.

Mitigations (actor-side).

- Cadence budgeting (caps on c_t), portfolio of channels (poly-platform), off-platform assets (email, events).
- Rotate roles: research, bridge, compliance, field ops—so the same people don't absorb all enforcement risk.
- Tie credibility to *verified outcomes* to counteract the engagement-only treadmill.

Mitigations (platform/governance).

- Reduce appeal latency; provide stable policy exemplars.
- Offer "exploration lanes" where cross-silo trials face lower default damping if paired with verification artifacts.

- Support third-party measurement of R_cross_t / R_local_t for contested topics to calibrate containment without entrenching it.

Summary of phase paths.

- **Visibility-debt plateau → Platform Thermidor** (durable quarantine).
- **Thermidor → Misfire → Snapback** (brief cascade, then re-tightening).
- **Thermidor with burnout** (bridge attrition, increasing brittleness).
  The healthiest exit ramps increase h_t (bridges) and mu (conversion efficiency), so that even modest L_t produces durable, verifiable change—shrinking the visibility debt without inviting systemic risk.

5. Measurement and Indicators

5.1 Reach-suppression ratio (impressions per follower vs baseline)

Goal. Quantify distribution throttling independent of audience size.

Metric. For a post p from account a at time t:

- ipf_p = impressions_p / followers_{a,t}

Aggregate by topic/window W (median or trimmed mean across posts):

- IPF_target(W), IPF_baseline(W)

Reach-suppression ratio (RSR).

- RSR(W) = IPF_target(W) / IPF_baseline(W)

Baselines (pick at least two):

1. Same account, low-risk topics matched by hour-of-week and format.
2. Matched accounts (size, niche, language) on same topic window.
3. Pre/post within-account DiD around a policy change.

Controls. Two-way fixed effects (hour-of-week, week-of-year), format dummies (video/image/link/text), and geo. Report bootstrap CIs. If RSR $\ll 1$ with tight CI, containment is likely.

Red flags. Falling non-follower share at constant follower growth; RSR breaks specifically on flagged keywords; long recovery (hysteresis) after strikes.

5.2 Attention half-life outside the silo

Goal. How fast does non-follower attention decay once exposed?

Setup. Collect non-follower impressions by elapsed time i_out(Δt) after first exposure (bin by minutes/hours).

Exponential estimate (robust default):

- Fit i_out(Δt) ≈ i0 * exp(−Δt / tau).
- Half-life t_half = tau * ln(2).

Heavy-tail alternative (no parametric fit):

- t_half = smallest T with cumulative_out(T) ≥ 0.5 * total_out_7d.

Diagnostics. Compare t_half_out vs t_half_in (followers). Containment shows short t_half_out with normal t_half_in. Track drift of t_half_out around policy toggles; a structural drop signals Platform Thermidor.

Bias controls. Normalize for content length/format; exclude paid boosts; stratify by entry channel (recommendation vs search vs share).

5.3 Cross-venue pickup and policy-translation ratios

Goal. Separate heat (coverage) from consequence (rules, budgets, services).

Windows. Choose k in {3d, 7d, 30d} from initial surge.

Cross-venue pickup ratio (CVPR_k).

- CVPR_k = legacy_mentions_k / (R_cross_k / 100k)

("Legacy mentions" = unique stories or segments in accredited outlets; dedupe syndications.)

Policy translation ratio (PTR_k).

- PTR_k = policy_artifacts_k / legacy_mentions_k

("Policy artifacts" = bill filings, rulemakings, hearings placed on calendar, budget lines, enforcement actions, FOIA wins delivered, signed MOUs; pre-register inclusion rules.)

Latencies.

- T_pickup = median lag post→first legacy mention.
- T_policy = median lag post→first artifact.

Benchmarks. In containment equilibria expect low CVPR_k, very low PTR_k, long T_pickup, and right-censored T_policy. Misfire/cascade moments show step-ups in both ratios with shorter latencies.

5.4 Search/sentiment deltas among non-followers

Goal. Detect belief/interest movement outside the silo.

Non-follower proxy cohorts.

- Geography without audience concentration ("cold" geos).
- Demographic slices under-represented in the follower graph.
- Lookalike audiences with explicit follower exclusion (where available).

Search delta (per cohort c).

- ΔSearch_c = Index_topic_c,postwindow − Index_topic_c,baselinewindow

Normalize by a placebo topic to get DiD:

- ΔΔSearch_c = (Index_topic_c,post − Index_topic_c,base) − (Index_placebo_c,post − Index_placebo_c,base)

Sentiment delta (per cohort c).

- Collect cross-platform public text outside follower graph; score with a calibrated, topic-specific model; report:
    - share_pos_c, share_neg_c, net_sent_c = share_pos_c − share_neg_c
    - Δnet_sent_c = net_sent_c,post − net_sent_c,base

Integrity checks.

- Pre-trend tests on placebo topics.
- Common-shock controls (major news unrelated to the silo).

- Leakage test: if ΔΔSearch rises only in follower-dense geos, you're measuring in-silo echoes, not public movement.

Minimal data schema (to make this reproducible)

- posts.csv: post_id, acct_id, ts, topic_tag, format, followers_t, impressions_followers, impressions_nonfollowers, strikes_count
- exposure_timeseries.parq: post_id, bucket_min, nonfollower_impressions
- media_mentions.csv: post_id, outlet_id, ts, url, dedupe_cluster_id
- policy_artifacts.csv: post_id, artifact_type, jurisdiction, ts, url
- search_index.csv: region, topic, day, index_value, is_placebo
- sentiment_samples.parq: text_id, region, day, follower_flag, score

---

Reading the dashboard at a glance

- **RSR < 0.4** with tight CIs → probable containment.
- **t_half_out** short and falling while **t_half_in** stable → Platform Thermidor.
- **CVPR_k and PTR_k** near zero across cycles → visibility debt.
- **ΔΔSearch_c and Δnet_sent_c** > 0 outside follower geos → real spillover; watch for cascade conditions next.

6. Strategies for Siloed Actors

6.1 Poly-platform routing and redundancy

Goal. Keep messages alive when any single platform raises s_t (suppression). Reduce single-point-of-failure and preserve R_cross_t via multiple discovery surfaces.

Routing plan (design once, reuse).

- **Own base**: canonical post on a site you control (stable URL, canonical tag, RSS/Atom). Mirror as a PDF/data pack (SHA256 + timestamp) for integrity.

- **Syndicate outward**: short forms to fast feeds (X, Threads), visuals to Reels/Shorts/TikTok, long form to Substack/Medium, discussion to Reddit/Forums, evidence to GitHub/OSF/Zenodo, audio cut to podcast feed.
- **Queue & DAG**: content goes to a publish queue; edges define cross-posts and retries. If rec=0 (no recommendations) is detected for node A, retry on B,C with format changes (clip, quote-card, carousel).
- **Email fallbacks**: newsletter + segmented mailing lists (press, officials, partners). Email is outside recommender throttles.

Operational hardening.

- Access: 2FA, credential rotation, role-based accounts; at least 2 admins per channel (bus-factor ≥ 2).
- Backups: export analytics weekly; keep local copies of posts and comments for audit.
- Failover drills: monthly simulation of an account lock—how fast until mirrors are live?

Metrics.

- **Per-platform RSR**: IPF_target / IPF_baseline (Section 5.1). Track dispersion, not just means.
- **Survival rate**: share of posts with non-zero non-follower impressions within 24h.
- **Cross-surface elasticity**: d(R_cross_total)/d(s_platform_i) estimated around outages.

6.2 Link laundering vs credibility loss: a tradeoff

Definition. “Link laundering” = routing the same claim through a different domain/brand (e.g., partner NGO, local paper, neutral repository) to escape topic/account dampers. It can raise R_cross_t short-run but risks trust if overused.

When it helps.

- Platform dampers key off **source domain** or **account trust**.
- Neutral wrappers (local press, civic orgs, preprint servers) are whitelisted and gain recommendability.

Risks.

- Audience senses indirection → **credibility penalty**.
- Partners burn goodwill if they front for overt advocacy.
- Moderation systems update; short-term lift decays.

Decision frame (simple).

- Let E_pickup = expected delta in legacy pickup from laundering.
- Let C_trust = estimated trust loss (survey or A/B on willingness to share).
- Proceed if E_pickup - C_trust > theta (theta = minimum acceptable net).

Guardrails.

- Disclose provenance inside the artifact ("Produced by X; hosted via Y for archiving").
- Prefer **substantive wrappers** (methods, data) over cosmetic mirrors.
- Use **few, durable venues**; avoid throwaway link shorteners.
- Periodically publish a **provenance log** to rebuild trust.

Metrics.

- Pickup elasticity: d(L_t)/d(launder_flag).
- Trust delta: change in "would share with a colleague" or "credible source" scores (panel).

6.3 "Proof-of-outcome" activism: FOIA wins, service delivery, verifiable impacts

Principle. Shrink visibility debt by producing artifacts that move tau_t even when L_t is small.

Tactics.

- **FOIA pipeline.** Maintain a tracker: request_id, filed_on, statutory_due, status, pages_released, appeal, key findings, posted_at. Publish releases with diff notes and citations that journalists can lift verbatim.
- **Service delivery.** Offer concrete help aligned with the claim (e.g., form-filling clinics, hotline hours, referral maps, small grants). Log "served_n", "median wait", "resolution_rate".

- **Rule/budget hooks.** Always attach a specific lever: docket number, budget line, meeting date, draft text. Provide "paste-ready" paragraph edits and a contact list.
- **Outcome ledger.** A public changelog (ledger_id, date, unit_changed, before_text, after_text, proof_link, hash). Make it boring and audit-friendly.

Templates (one-liners).

- **FOIA subject line**: "FOIA Request: [Program/Office] — [Document type] — [Date range]".
- **Outcome post**: "Outcome: [unit] changed [X→Y]. Proof: [URL]. Next lever: [docket/budget]."

Quality controls.

- Timestamp + hash every artifact; archive snapshots.
- Adopt a **retraction protocol** (clear, fast, loud) to protect long-run q_t.

Metrics.

- **PTR_k** (policy translation ratio) up over rolling windows.
- **Latency medians**: T_pickup, T_policy (Section 5.3) trending down.
- **Cost per win**: staff_hours / artifact; use to prioritize replicable plays.

6.4 Bridge nodes: cultivating cross-silo intermediaries

Goal. Raise h_t (path entropy) and x_t (bridge quality) so R_cross_t improves without waiting on s_t to fall.

Who bridges.

- Local reporters, trade press, beat bloggers.
- Domain pros with mixed audiences (engineers, clinicians, teachers).
- Faith/civic leaders, union stewards, small-biz associations.
- "Adjacent" creators respected by out-groups.

Finding bridges (practical heuristics).

- **Overlap scan**: accounts followed by both your followers and target out-group (Jaccard > tau).

- **Cite graph**: who mainstream outlets quote on adjacent topics?
- **Warm intros**: map 2-hop ties via LinkedIn/Twitter lists; ask for 15-minute "method checks," not coverage.

Bridge brief (1-pager skeleton).

1. *What is new* (2–3 bullets, no jargon).
2. *Why it matters* (one concrete stake: dollars, time, safety).
3. *Evidence pack* (dataset link, methods note, 2 visuals).
4. *Verifiable hook* (docket/budget/event date).
5. *Who to call* (two sources with phones/email).
6. *Embargo & usage* (clear terms).

Working norms.

- No dunking, no gotchas in bridge channels.
- Offer co-credit for follow-ups; share preprints and raw data.
- Keep "asks" small and specific; follow with an outcome you can thank them for.

Protect the bridges.

- "Firebreak" vocabulary for sensitive terms; provide context blocks they can copy.
- Let them opt into **tiered briefings** (monthly zoom, embargoed notes).
- Do not route every hot take through the same person.

Metrics.

- **Bridge count**: active intermediaries per quarter.
- **Hit rate**: share of briefs that produce any non-follower exposure.
- **Lift**: delta in R_cross_t / R_local_t in the 7d after a bridge placement.
- **Downstream**: increases in L_t and tau_t attributable to bridge-originated coverage (match via UTM/citation).

Putting it together (weekly cadence)

- **Mon**: Publish canonical post + data pack; queue cross-posts; send two tailored bridge briefs.
- **Tue–Wed**: Service delivery block (clinic/hotline); file or follow FOIA(s).
- **Thu**: Outcome ledger update; "paste-ready" paragraph for officials.
- **Fri**: Metrics review (RSR, t_half_out, CVPR_k, PTR_k); adjust routing and bridge list.
- **Monthly**: Failover drill; re-score bridges; retire plays with poor cost per win.

Exit criterion from visibility debt.
If PTR_k rises steadily while RSR remains low, keep investing in bridges + outcomes. If PTR_k stagnates and burnout signals rise, pause rhetoric volume, consolidate proofs, and re-center on one lever where mu (conversion efficiency) is demonstrably > 0.

7. Governance and Design Options

7.1 Auditable ranking signals and contestable downranking

Objective. Make distribution throttles legible and appealable without exposing the full model to gaming.

Mechanisms.

- **Why-This-Reach (WTR) card.** Per-post, per-audience explanation showing coarse weights for major signal families: freshness, user trust, source diversity, prior corrections, policy flags, and "context added" events. Show bins, not raw scores (e.g., low/med/high).
- **Ranking Transparency Ledger (RTL).** Signed, append-only logs that record the presence of downrank classes applied to a post (e.g., "health-claims/low-evidence"), with time stamps and decay horizons. Creators can export their own RTL slices.
- **Decision codes and SLAs.** Every throttle/strike maps to a stable codebook (D001 misinformation, D142 brigading). Appeals have response-time targets (e.g., T_1 = 24h triage, T_2 = 7d final).
- **Contestability API.** Structured appeals that attach counter-evidence (citations, datasets, corrections). On success, systems must (i) remove the flag, (ii) issue a repair boost for a bounded window, and (iii) log the reversal in RTL.

- **Independent auditors.** Accredited third parties can sample anonymized RTL records with differential privacy to publish quarterly reports (topic-level RSR distributions, false-positive rates).

Safeguards against gaming.

- Coarse quantization of signals, rate-limited exports, randomized audit sampling.
- Post-hoc, not pre-publish, disclosures; delayed release of some flags during live harm events.

Metrics.

- Appeal latency (median, p90), reversal rate, post-reversal reach restoration, false-positive/false-negative estimates, creator trust surveys.

7.2 Frictions that reward verification over virality

Objective. Rebalance incentives: make it easier to spread *verified* claims than unverified hype.

Mechanisms.

- **Verify-to-Boost.** Posts with linked evidence packs (dataset, method note, source citations) qualify for a small cross-silo lift; posts lacking evidence incur a mild default damper outside followers. Evidence can be attached within a grace window (e.g., 24h) to remove the damper.
- **Read-Before-Reshare.** For long articles and studies, require an open + dwell or a summary quiz before cross-silo resharing triggers. Inside the silo, standard resharing remains.
- **Cooling-off prompts.** When a post's out-of-silo resharing velocity exceeds a threshold, add a "context preview" interstitial with links to counter-evidence and prior corrections.
- **Credibility credits.** Maintain an account-level "outcome credit" that increases with retractions issued, FOIA/artifact uploads, and successful corrections; decreases with upheld strikes. Credits modulate default recommendability for adjacent topics.
- **Correction dividend.** Rapid, prominent corrections receive a short-term distribution bump on the corrected item and a small, time-bounded lift to adjacent posts.

Metrics.

- Differential reach for verified vs unverified posts, correction latency, correction uptake rate, change in misinformation complaint rates, user completion of context prompts, cost of false frictions (drop-off on legitimate news).

7.3 Pluralizing cues: diversified recommendation neighborhoods

Objective. Reduce echo intensity by ensuring exposure to heterogeneous but credible sources.

Mechanisms.

- **Neighborhood diversity constraint.** For non-follower recommendations, enforce a minimum path-entropy target: each session includes K items sampled from distinct clusters with bounded ideological similarity. Implement via reweighted random walks or determinantal point processes to avoid near-duplicates.
- **Qualified heterodoxy.** Diversity draws only from eligibility sets that clear harm and quality screens (no "false balance" on dangerous content). Eligibility criteria and appeals mirror Section 7.1.
- **Cue providers registry.** Curate and publish a rotating list of local journalism outlets, trade press, civic orgs, and cross-community creators. Provide a submission and audit route; show inclusion/removal rationales.
- **User controls with guardrails.** Let users tilt their diversity dial (e.g., "standard", "broader", "minimal"), with periodic nudges back toward "standard." Show a small session-level explainer: "Included viewpoints from X, Y for context."

Metrics.

- Cross-silo exposure ratio, average time-to-first opposing view, complaint/abandon rates, retention among users with elevated diversity, effect on non-follower t_half and on RSR dispersion.

Risks and mitigations.

- **Perceived forced exposure.** Offer granular controls and clear explanations.
- **Attention tax on routine users.** Apply constraints only to cross-silo recommendations, not follower feeds.

7.4 Safe harbors for cross-silo deliberation

Objective. Create spaces optimized for good-faith contestation, protected from default containment dynamics.

Mechanisms.

- **Deliberation rooms.** Invitation-based or topic-bounded spaces with a posted charter, facilitation, and structured templates (claim, warrant, source, counter-argument). Participation confers a temporary immunity from certain downrank classes provided rules are followed.
- **Argument mapping and adversarial collaboration.** Built-in tools to create issue maps; periodic "steelman" rounds where participants must articulate opponents' best case; paired teams tasked with joint statements or decision memos.
- **Provenance and transcripts.** Auto-captured, tamper-evident logs (hash-stamped) with privacy-respecting redactions. Outputs (executive summaries, data links) are publishable artifacts eligible for Verify-to-Boost.
- **Moderator commons.** Trained, rotating facilitators drawn from multiple communities; shared playbooks; transparent escalation paths; sanctions for brigading.
- **No-penalty participation.** Contributions in good standing do not carry topic-risk penalties to the contributor's broader account score.

Policy supports.

- **Regulatory safe harbor.** Compliance with an audited deliberation charter confers liability protections similar to documented trust-and-safety programs.
- **Funding rails.** Creator payments or grants tied to quality signals (peer endorsements across camps, method checks, external uptake in policy or practice).

Metrics.

- Pre/post belief movement (panel), cross-camp reciprocity rates, artifact reuse by media/policymakers (L_t, tau_t upticks), civility/turn-taking measures, churn of bridge participants, incidence of brigading.

Implementation notes and trade-offs

- **Privacy vs transparency.** Use coarse bins, randomized audits, and differential privacy. Reveal enough to contest decisions, not enough to fully reverse-engineer ranking.

- **Friction costs.** Measure drop-offs; tune prompts to appear only when diffusion risk is high (velocity, novelty, topic).
- **Abuse risks.** Prevent coordinated laundering by weighting credibility credits toward third-party verified outcomes and penalizing synthetic networks.
- **Competition and portability.** Expose WTR cards, RTL slices, and credibility credits via standardized export so creators can port reputation across platforms.

Success criterion. Over rolling quarters: (i) RSR gaps narrow for verified content without overall surge in harm complaints, (ii) t_half_out lengthens for substantive items, (iii) CVPR_k and PTR_k rise on topics with outcome artifacts, and (iv) bridge survival rates improve—evidence that attention is reallocating from spectacle to verifiable signal.

8. Tests, Competing Explanations, and Falsifiers

8.1 What empirical patterns would disconfirm "quarantine"

Our "quarantine" claim predicts: R_local >> R_cross, short t_half_out, low CVPR_k and PTR_k, bridge attrition, and sensitivity of cross-silo reach to platform suppression s_t. The model is wrong if we observe durable, repeated deviations that survive basic controls.

A. Distribution & exposure falsifiers

- **No cross/follower gap:** For flagged topics, RSR ≈ 1 (impressions-per-follower at parity with baseline) for ≥ 8 consecutive weeks after controlling for hour/format/seasonality.
  Interpretation: downranking not binding.
- **Long non-follower dwell:** t_half_out comparable to t_half_in across multiple campaigns, independent of strikes or policy windows.
  Interpretation: no systematic decay outside the silo.
- **Path entropy growth under suppression:** H_path (or normalized h_t) rises while s_t (policy flags/strike rates) is stable or increasing.
  Interpretation: cross-silo connectivity is expanding despite supposed containment.

B. Media & policy falsifiers

- **High consequence without pickup:** PTR_30d remains high (e.g., ≥ 0.25) while L_t (legacy pickup) stays near zero across jurisdictions lacking strong offline capacity.
  Interpretation: policy change is occurring without E-mediated cues or offline infrastructure—contrary to our pathways.

- **Pickup insensitive to risk:** L_t shows no negative association with lambda_t (platform/advertiser/regulatory risk proxies) in time-series or panel models.
  Interpretation: elite amplification isn't gated by platform risk climate.

C. Platform-shock falsifiers (quasi-experiments)

- **No response to moderation toggles:** Around documented policy changes/outages, DiD estimates show R_cross and CVPR_k flat for treated topics/accounts vs matched controls.
  Interpretation: cross-silo reach is not mediated by ranking containment.
- **Appeal reversals without repair effects:** When flags are overturned, post-reversal reach restoration ≈ 0 across many cases.
  Interpretation: containment wasn't the constraining margin.

D. Bridge ecosystem falsifiers

- **Bridge resilience:** Bridge nodes' ROI (non-follower impressions per post) is equal/higher than in-silo specialists over long windows with steady s_t.
  Interpretation: quarantine is not eroding cross-silo edges.

Quick falsifier checklist (any two, sustained ≥ 2 months, with controls):
(i) RSR_target ≈ RSR_baseline and t_half_out ≈ t_half_in; (ii) high PTR_k without L_t in low-capacity locales; (iii) H_path ↑ while s_t steady/high.

8.2 Distinguishing resource poverty from algorithmic suppression

Low cross-silo reach could stem from *binding resource/skill constraints* rather than ranking containment. We separate them with interventions and natural experiments that isolate the margin of constraint.

1) Randomized resource shocks (field RCT)

- **Design:** Randomly assign micro-grants (e.g., ad credits, editing time, data viz help), professionalized packaging (evidence packs), and paid bridge outreach to a treatment set; control group unchanged.
- **Interpretation:**
  - *Resource-binding world:* R_cross, CVPR_k, and PTR_k rise materially in treatment.
  - *Suppression-binding world:* treatment lifts quality/format but leaves R_cross flat; RSR remains low.

2) Cross-platform elasticity (difference-in-differences)

- **Design:** Simultaneously post standardized content on platforms with known differences in moderation posture; compute d R_cross_total / d s_platform.
- **Interpretation:**
    - Large cross-platform dispersion aligned with s_platform → suppression signal.
    - Uniformly low performance → content/fit/resources likely binding.

3) Time-randomization & saturation tests

- **Design:** Randomize posting slots within creator's peak windows; add controlled reposts. If results hinge on resource-driven "bad timing," performance should improve with optimization.
- **Interpretation:**
    - *Improves with timing/saturation:* resource/ops deficit.
    - *Insensitive to timing, still low out-of-silo:* suppression.

4) Whitelisted partner replication ("source-cred" instrument)

- **Design:** Place identical artifacts with a high-trust, whitelisted outlet (local press, trade journal) and compare non-follower exposure to the original post.
- **Interpretation:**
    - Strong lift via partner while original stays damped → source/account gatekeeping (suppression).
    - Both low → demand/fit/resources.

5) Paid reach sanity check (limited, transparent)

- **Design:** Small, targeted paid placements to non-followers (no behavioral targeting beyond geography/topic).
- **Interpretation:**
    - Paid reach yields normal click-through and downstream L_t/PTR_k → organic suppression likely binding.
    - Paid reach underperforms benchmarks → message/product-market mismatch or credibility/resource issues.

6) Quality & clarity A/B (content-side instrument)

- **Design:** Randomized rewrites that increase clarity/evidence density without changing claim substance; pre-registered scoring rubric.
- **Interpretation:**
    - Large lifts in R_cross with clarity → quality was the margin.
    - Flat R_cross despite better quality and stable R_local → suppression.

7) Appeals latency & repair responsiveness

- **Design:** Track appeal outcomes, time to reversal, and post-reversal restoration.
- **Interpretation:**
    - Fast reversals + restored reach → working contestability; resource constraints may dominate.
    - Slow/no reversals + no restoration → persistent containment beyond resources.

8) Offline conversion probe (μ vs s_t)

- **Design:** Hold online cadence steady; invest in offline capacity (clinics, FOIA pipeline).
- **Interpretation:**
    - PTR_k rises while R_cross stays low → resource/infrastructure was binding (our model allows μ↑ to raise τ even under suppression).
    - PTR_k stuck at ~0 despite new offline capacity → broader demand/credibility problem or hard suppression of pickup channels.

Minimal analysis recipes

- **Panel FE model (account×week):**
  $R_cross_{it} \sim \beta1 * s_{it} + \beta2 * quality_{it} + \beta3 * timing_{it} + \alpha_i + \delta_t + \varepsilon_{it}$
  Suppression signal if $\beta1 << 0$ after controls; resource signal if β2, β3 drive fit.
- **Synthetic control around policy toggles:** build donor pool from unflagged topics to estimate counterfactual R_cross.
- **Event studies:** plot ±30d around strikes, appeal reversals, ranking policy updates; look for sharp breaks in RSR, t_half_out, CVPR_k.

Decision table (diagnosis → action)

Finding	Likely cause	Priority action
RSR low, insensitive to timing/quality; partner replication lifts	Suppression/gatekeeping	Appeals; provenance packs; bridge via whitelisted outlets; poly-platform routing
RSR improves with packaging/timing; CVPR_k up with grants	Resource/ops	Keep grants; standardize templates; hire ops/editor
Paid reach normal but organic flat	Algorithmic throttle	Contestability + diversify discovery (search/newsletters)
Both paid and organic weak; sentiment deltas flat	Demand/message fit	Revise audience/offer; test adjacent frames; add service delivery
PTR_k rises with offline investment while R_cross flat	Offline μ was binding	Scale clinics/FOIA; keep modest online cadence

If multiple tests point away from suppression, the "quarantine" story should be down-weighted in favor of resource, quality, or demand explanations. If multiple tests point toward suppression, the model is strengthened—and so is the case for the governance levers in Section 7.

9. Limitations and Ethical Considerations

9.1 Privacy, measurement bias, and adversarial inference

Privacy risks. Our indicators (RSR, t_half_out, CVPR_k, PTR_k, ΔΔSearch) aggregate user behavior that can, in small cohorts or rare-topic locales, enable re-identification or sensitive-inference (political views, health status).

- **Minimize & bound.** Collect only derived metrics; avoid storing raw text/handles. Apply retention limits (e.g., rollups kept 90 days; raw events discarded in ≤ 7 days).
- **Aggregation discipline.** Require cohort size ≥ k (e.g., k ≥ 100) and geography ≥ region-level; suppress small cells and outliers.
- **Noise & budgets.** Add calibrated noise to public dashboards (ε-bounded differential privacy). Publish an annual "privacy budget" and depletion logs.

- **Provenance hashing.** Hash artifact links and store hashes, not URLs, when feasible; prevent linking to personal identifiers.
- **Access tiers.** Internal fine-grained telemetry in secure enclaves; external releases coarse and DP-protected; researcher access via vetted data use agreements.

Measurement bias. Platform metrics are not ground truth and can encode sampling and survivorship bias.

- **Selection effects.** Silo members differ from the public; correct with post-stratification weights against census-like priors where available.
- **Confounding & seasonality.** Control for hour-of-week, format, language, and news-cycle shocks; pre-register models for DiD and event studies; report robustness to alternative specifications.
- **Leakage between measures.** CVPR_k may be inflated by syndication; dedupe via cluster IDs. PTR_k can be overstated by symbolic "policy artifacts" (e.g., placeholder hearings); pre-define inclusion/exclusion rules.
- **Missing data.** Document appeal outcomes, demonetization flags, and outages; mark analyses "incomplete" when these covariates are unknown.

Adversarial inference & gaming.

- **Deanonymization/linkage.** Cross-platform linkage or small-cell reports can expose identities; enforce k-anonymity and DP before public release; rate-limit API exports.
- **Metric gaming.** Goodhart risk: if RSR or PTR_k become targets, actors may spam bridges or fabricate "policy artifacts." Mitigate with audit sampling, artifact verification (hashes, docket checks), and penalties for synthetic networks.
- **Model inversion & membership inference.** Do not release raw embeddings; use privacy-preserving sketches for path-entropy estimates; randomize audit samples.
- **Brigading & poisoning.** Monitor for coordinated posting/signal inflation; use rolling anomaly detectors and holdout evaluations that exclude suspicious clusters.

Governance of the measurement program.

- Treat this as human-subjects research: independent ethics review, public documentation of instruments, red-team drills for re-identification, and a public incident log with remediation timelines.

9.2 Risks of chilling legitimate dissent

Our levers (Verify-to-Boost, cooling-off prompts, WTR/RTL transparency, safe harbors) can inadvertently dampen urgent speech, disadvantage low-resource voices, or entrench viewpoint bias.

Where chilling can occur.

- **Friction asymmetry.** Read-before-reshare or context prompts raise costs for time-sensitive whistleblowing more than for well-staffed outlets.
- **Proof burdens.** Verify-to-Boost may privilege groups able to package data/methods; marginalized speakers could be de-amplified despite accuracy.
- **Appeal friction.** Slow or complex appeals reduce contestability for small actors; visibility lost during delay is rarely recovered socially even if repaired algorithmically.
- **Viewpoint signals.** Topic-risk buckets can proxy ideology; if not tightly scoped, they become de facto viewpoint filters.

Guardrails and tests.

- **Minimum-necessary friction.** Trigger prompts only under high diffusion risk (velocity × novelty × topic). Provide "emergency bypass" with after-the-fact review for whistleblowing and crisis alerts.
- **Equity audits.** Run quarterly disparate-impact tests by language, region, resource level (solo creators vs institutions). If RSR_gap(group) > θ after quality/timing controls, mandate policy review.
- **Assisted verification.** Offer low-bar "evidence packs" (templated affidavits, timestamped photos, auto-generated method notes) and micro-grants for packaging; recognize trustworthy intermediaries (legal aid, local press) as acceptable verifiers.
- **Appeal SLAs & repair.** Set and publish appeal time targets; when reversals occur, apply time-bounded "repair boosts" and annotate WTR cards so audiences see the correction.
- **Viewpoint neutrality.** Define risk classes around *behavior and harm* (brigading, doxxing, fabricated evidence), not ideology. Require dual-key sign-off for any addition to risk taxonomies; publish examples and counter-examples.
- **Sunset & review.** Every new friction or downrank rule gets an automatic sunset (e.g., 6 months) unless re-authorized after impact review with public comment.

Participatory design.

- Include creators, journalists, civil rights groups, and subject-matter experts in policy iteration. Provide localized language support and plain-English policy explanations.

Outcome-over-rhetoric bias.

- Where possible, bias lifts toward *verifiable outcomes* (FOIA releases, corrections issued, service delivered) rather than toward specific frames. This rewards truth-seeking behavior without privileging any ideology.

Monitoring chilled-speech signals.

- Track spikes in "self-censorship" survey items, anomalous drops in posting from high-risk cohorts, and increased off-platform migration following policy changes. Treat these as leading indicators requiring rapid policy tuning.

Bottom line. The same tools that can reduce systemic risk can also mute legitimate dissent if left unchecked. Make thresholds explicit, appeals fast, evidence assistance real, and reviews participatory—so that the system constrains *harmful diffusion* without penalizing *truthful challenge*.

Appendices (Operationalization)

A. Data schema and metrics cookbook

A.1 Minimal relational schema (3NF, analytics-friendly)

```
-- 1) Accounts and posts
CREATE TABLE accounts (
  acct_id          TEXT PRIMARY KEY,
  platform         TEXT,        -- e.g., x, yt, ig, tt, reddit, web
  handle           TEXT,
  created_at       TIMESTAMP,
  is_org           BOOLEAN,
  topic_primary    TEXT         -- optional tag for stratification
);

CREATE TABLE posts (
  post_id          TEXT PRIMARY KEY,
  acct_id          TEXT REFERENCES accounts(acct_id),
  platform         TEXT,
  ts_published     TIMESTAMP,
  topic_tag        TEXT,        -- analyst-coded or model-coded
  format           TEXT,        -- text|image|video|link|carousel|longform
  url_hash         TEXT,        -- SHA256(url) or canonical artifact hash
  strike_count     INT DEFAULT 0, -- snapshot at publish time
  demonetized      BOOLEAN,
  policy_flags     TEXT[],      -- e.g., ['health-claims/low-evidence']
  appeal_status    TEXT         -- none|pending|won|lost
);
```

```
-- 2) Exposure & impressions (followers vs non-followers)
CREATE TABLE exposure_summary (
  post_id          TEXT REFERENCES posts(post_id),
  platform         TEXT,
  followers_at_post  INT,
  imps_followers     BIGINT,
  imps_nonfollowers  BIGINT,
  clicks_nonfollowers BIGINT,
  dwell_ms_nonfollowers BIGINT,
  PRIMARY KEY (post_id, platform)
);

-- Optional high-res time series (downsampled bins)
CREATE TABLE exposure_timeseries (
  post_id          TEXT,
  bucket_start      TIMESTAMP,   -- e.g., 5 or 15 min bins
  imps_nonfollowers  INT,
  clicks_nonfollowers INT,
  PRIMARY KEY (post_id, bucket_start)
);

-- 3) Media pickup & policy artifacts
CREATE TABLE media_mentions (
  mention_id        TEXT PRIMARY KEY,
  post_id          TEXT,       -- originating item (nullable if external)
```

```
  outlet_id         TEXT,
  ts_mentioned      TIMESTAMP,
  url_hash          TEXT,
  cluster_id        TEXT        -- for deduping syndication
);

CREATE TABLE policy_artifacts (
  artifact_id       TEXT PRIMARY KEY,
  post_id           TEXT,
  jurisdiction      TEXT,       -- e.g., US-CA, city:Boston
  artifact_type     TEXT,       -- bill|hearing|rulemaking|budget|FOIA|enforcement|MOU
  ts_opened         TIMESTAMP,
  ts_closed         TIMESTAMP,  -- nullable
  status            TEXT,       -- filed|calendar|passed|implemented|released
  url_hash          TEXT
);

-- 4) Search & sentiment (cohorts exclude followers)
CREATE TABLE search_index (
  day               DATE,
  region            TEXT,       -- DMA/state/country
  topic             TEXT,
  idx_value         NUMERIC,    -- normalized 0..100 or z-score
  is_placebo        BOOLEAN,
  PRIMARY KEY (day, region, topic)
);
```

```
CREATE TABLE sentiment_samples (
  sample_id         TEXT PRIMARY KEY,
  day               DATE,
  region            TEXT,
  follower_flag     BOOLEAN,     -- false => non-follower cohort
  topic             TEXT,
  score             NUMERIC      -- -1..+1 or 0..1
);
```

Integrity & privacy notes

- Store url_hash (SHA256) instead of raw URLs when possible.
- Enforce k-anonymity on public extracts (e.g., cohorts n >= 100).
- Keep raw event logs ≤ 7 days; aggregate to the above tables weekly.

A.2 Metric definitions (copy-safe, implementation-ready)

Let W be an analysis window (topic × time range × platform).

Impressions-per-follower (IPF)
IPF_post = imps_total / followers_at_post
IPF_target(W) = median(IPF_post over posts in target topic/window)
IPF_baseline(W) = median(IPF_post over matched baseline posts)

Reach-Suppression Ratio (RSR)
RSR(W) = IPF_target(W) / IPF_baseline(W)

Non-follower half-life (t_half_out)
Fit i_out(dt) ≈ i0 * exp(−dt / tau) using exposure_timeseries for non-followers.
t_half_out = tau * ln(2)
(Nonparametric fallback: minimal T s.t. cum_out(T) ≥ 0.5 * total_out_7d.)

Cross-venue Pickup Ratio (CVPR_k)
Let k ∈ {3d, 7d, 30d}.
CVPR_k = legacy_mentions_k / (R_cross_k / 100k)

where R_cross_k = sum(imps_nonfollowers over window k).
Deduplicate mentions via cluster_id.

Policy Translation Ratio (PTR_k)
PTR_k = policy_artifacts_k / legacy_mentions_k
Include only artifacts with status ∈ {calendar, passed, implemented, released}.

Path entropy proxy (h_t)
Estimate community exposure distribution P(c) from cross-silo non-follower cohorts;
H = −Σ_c P(c) * log P(c); normalize: h = H / log K (K = num communities).

Repair restoration (post-appeal)
Restoration = (IPF_post_after / IPF_baseline_matched) − (IPF_post_before / IPF_baseline_matched)

A.3 ETL & quality controls (checklist)

- Time normalization: convert all timestamps to UTC; store user local tz separately if needed.
- Seasonality controls: add hour-of-week FE; exclude major holidays unless explicitly analyzed.
- Format controls: include dummies for text/image/video/link; stratify when sample allows.
- Outlier handling: Winsorize top 1% of IPF and non-follower impressions; log-scale where needed.
- Dedupe: media cluster_id (URL similarity + near-duplicate text), policy artifacts by docket/bill number.

B. Simulation sketch for I–G_c–E–P dynamics

Goal: reproduce three regimes—(i) containment equilibrium, (ii) leakage, (iii) cascade + snapback.

B.1 State and parameters (discrete time, t = 1..T)

States:

- R_local_t, R_cross_t

- h_t in [0,1] (path entropy)
- s_t in [0,1] (suppression)
- L_t in [0,1] (legacy pickup)
- tau_t in [0,1] (policy translation)

Controls / shocks:

- x_t >= 0 (bridge quality)
- q_t in (0, +) (content quality/trust)
- lambda_t >= 0 (platform risk aversion)
- kappa_t >= 0 (elite pickup hunger)
- v_t in [0,1] (veto friction)
- X_t ∈ {0,1} (exogenous shock)

Constants:

- R0 (base potential reach), mu ∈ (0,1] (conversion efficiency), Theta (spill threshold)

B.2 Update equations (copy-safe)

1. R_cross_t = R0 * (1 - s_t) * x_t * f(h_t) * q_t, with f(h) = eps + (1 - eps)*h
2. L_t = sigmoid(a0 + a1*R_cross_t + a2*kappa_t + a3*X_t - a4*lambda_t)
3. tau_t = L_t * (1 - v_t) * mu
4. h_{t+1} = h_t + eta1*bridge_invest_t - eta2*silo_opt_t - eta3*s_t
5. s_{t+1} = s_t + gamma1*Risk_t - gamma2*Audit_t - gamma3*Compete_t
6. Risk_t = r0 + r1*R_cross_t + r2*policy_flags_t
7. R_local_{t+1} = R_local_t*(1 + rho_local*engagement_t - delta*burnout_t)

sigmoid(z) = 1/(1 + exp(−z)).
eps is a small constant (e.g., 0.05) to avoid collapse when h=0.

Snapback rule (platform Thermidor):
If R_cross_t > Theta and Risk_t high for 2 consecutive steps, then set s_{t+1} = min(1, s_t + s_snap).

B.3 Pseudocode (language-agnostic)

```
init s=0.5, h=0.25, x=0.4, q=1.0, lambda=0.8, kappa=0.3, v=0.6

for t in 1..T:

  R_cross = R0 * (1 - s) * x * (eps + (1-eps)*h) * q

  L = sigmoid(a0 + a1*R_cross + a2*kappa + a3*X[t] - a4*lambda)

  tau = L * (1 - v) * mu


  Risk = r0 + r1*R_cross + r2*policy_flags[t]

  s_next = clamp( s + gamma1*Risk - gamma2*Audit[t] - gamma3*Compete[t], 0, 1 )


  h_next = clamp( h + eta1*bridge_invest[t] - eta2*silo_opt[t] - eta3*s, 0, 1 )


  if R_cross > Theta and Risk > Risk_hi for 2 steps:

     s_next = clamp( s + s_snap, 0, 1 )


  record {R_cross, L, tau, s, h}


  s = s_next

  h = h_next

end
```

B.4 Suggested parameter sets

- **Containment equilibrium (CE):**
  s=0.7, h=0.2, x=0.3, lambda=1.0, kappa=0.2, v=0.7, mu=0.4, Theta high
  Expect: low R_cross, L near 0, tau ~ 0.
- **Leakage window (LR):**
  temporary: s ↓ by 0.2 or h ↑ by 0.3 (bridge_invest spike), or X_t=1 with kappa ↑.
  Expect: R_cross bump, modest L, small tau.

- **Cascade + snapback (CC):**
  s moderate (0.4), h moderate (0.5), kappa high (0.8), v low (0.3), Theta moderate.
  Expect: R_eff > 1 for several steps, then s_snap = +0.25 raises s and ends cascade.

B.5 Plots to inspect

- Trajectories: R_cross, L, tau, s, h vs t.
- Phase plane: (h, s) with arrows showing drift.
- R_eff(t) crossings vs 1 (early-warning).

C. Survey modules for private/public belief gaps

Purpose: estimate preference falsification and cue-dependence outside the silo.

C.1 Core constructs

- **Public stance (PS):** what respondents say they believe/do when identity is visible.
- **Private stance (PrS):** what they believe/do under anonymity or indirect elicitation.
- **Cue dependence (CD):** reliance on elite/platform cues for updates.

C.2 Instruments (modular)

(i) Direct vs anonymized response (split-sample)

- Group A (identified): standard Likert items (1–5).
- Group B (anonymous box or randomized-response): same items.

Example item pairs (topic-neutral placeholders):

- “Policy X improves everyday life.” [1..5] (public)
- “Privately, I believe Policy X improves everyday life.” [1..5] (anonymous)

Gap metric: Gap_i = mean(PrS_i) − mean(PS_i) across matched samples.
Report CIs; pre-register primary items.

(ii) List experiment (unmatched count technique)

- Control list: 4 innocuous statements; Treatment list: same + sensitive statement.
- Ask: “How many of these do you agree with?”
- Difference in means estimates prevalence of the sensitive belief.

(iii) Endorsement experiment

- Randomize the attributed source (legacy elite vs micro-elite vs neutral) for the same policy vignette.
- Outcome: support score.
- CD proxy: delta in support when source flips.

(iv) Randomized response (RR) for stigmatized behavior

- Ask yes/no; instruct respondents to follow a private coin flip: heads = answer "yes", tails = answer truthfully.
- Estimate true prevalence with known p(heads).

(v) Perceived climate & self-silencing scale

- Items (1..5):
    - "People like me can share views on [platform] without risk."
    - "I avoid posting to prevent conflict with friends/colleagues."
    - "I wait for trusted figures before forming an opinion on new issues."
- Construct indices: Climate (safety), Self-silence, Cue-reliance.

C.3 Sampling & fielding

- Target N per arm: 1,200–1,500 for 80% power on 3–5pp effects.
- Stratify by region, age, education, media diet.
- Exclude followers of focal silo accounts from the "public" cohort to avoid contamination; alternatively, tag and control.

C.4 Analysis plan (pre-register)

- Compute Gap_i per topic; test H0: Gap_i = 0 with Welch t-tests.
- Difference-in-differences for events (before/after platform policy change).
- Multilevel models with random intercepts by region and media diet.
- Multiple-testing control (Benjamini–Hochberg) across items.

C.5 Example item bank (rotate to reduce demand effects)

Public/Private (paired):

- “I support increasing/decreasing funding for Program Y.”
- “I would share an article supporting Program Y.”
- “I would sign my name to a letter about Program Y.”

Endorsement sources (randomized label in brackets):

- “[Source: local newspaper / national cable host / prominent creator / nonpartisan NGO] says Program Y cut wait times by 30%. How credible is this?” [1..5]
- “Given the above, would you support trying Program Y in your area?” [1..5]

Perceived climate:

- “Posting about Program Y could harm my job or relationships.” [1..5]
- “On [platform], important corrections are easy to find.” [1..5]

Cue-reliance:

- “I often wait for [source category] before sharing an opinion.” [1..5]

C.6 Ethics & privacy

- Minimal PII; anonymous arms store no linkage keys.
- Consent language clearly explains RR/list-experiment protections.
- Pre-commit to destroying raw IPs within 7 days; aggregate to region.
- IRB/independent ethics review; publish instrument and analysis code.

Deliverables checklist (to reproduce the paper’s claims)

- **Data:** CSV/Parquet for all tables above, with documentation.
- **Code:** Notebooks/scripts to compute RSR, t_half_out, CVPR_k, PTR_k, h_t, restoration.
- **Sim:** A parameterized simulation script with presets CE/LR/CC.
- **Survey:** Questionnaire, sampling frame, pre-registration, analysis plan.
- **Hashes:** SHA256 for all public artifacts; provenance log with timestamps.

References

1. Gramsci, A. (1971). *Selections from the Prison Notebooks* (Q. Hoare & G. N. Smith, Eds.). International Publishers.
2. Kuran, T. (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.
3. Noelle-Neumann, E. (1984). *The Spiral of Silence: Public Opinion—Our Social Skin*. University of Chicago Press. (Orig. 1974.)
4. Weber, M. (1978). *Economy and Society: An Outline of Interpretive Sociology* (G. Roth & C. Wittich, Eds.). University of California Press.
5. Turchin, P. (2016). *Ages of Discord: A Structural-Demographic Analysis of American History*. Beresta Books.
6. Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.
7. Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press.
8. Bail, C. A. (2021). *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press.
9. Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
10. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
11. Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132.
12. Guess, A., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 U.S. election and its aftermath. *Nature Human Behaviour*, 4, 472–480.
13. Bail, C. A., Argyle, L. P., Brown, T. W., et al. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216–9221.
14. King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows criticism but silences collective action. *American Political Science Review*, 107(2), 326–343.

15. Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., et al. (2017). “You Can’t Stay Here”: The efficacy of Reddit’s 2015 ban examined. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 1(CSCW), 31:1–31:22.

16. Ribeiro, M. H., Ottoni, R., West, R., et al. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, 14, 438–449. (Preprint 2019: arXiv:1908.08313.)

17. Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1), 76–91.

18. Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329(5996), 1194–1197.

19. Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9), 5766–5771.

20. Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303–330.

21. Papacharissi, Z. (2015). *Affective Publics: Sentiment, Technology, and Politics*. Oxford University Press.

22. Roberts, M. E. (2018). *Censored: Distraction and Diversion Inside China’s Great Firewall*. Princeton University Press.

23. Guess, A., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2021). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 118(23), e2019000118.

24. Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, 5(2), 1–13.

25. Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of Blocklists. *ACM Transactions on Computer-Human Interaction*, 25(2), 12:1–12:33.

Note: Items 23–25 provide empirical background on interventions, moderation labor, and media literacy—useful for Sections 6–7 (strategies and governance). If you prefer a shorter list, we can prune to the core theoretical and measurement anchors (1–16).